世界のデータアーカイブの状況

データアーカイブ会議@シドニー（2019.8）＋個別調査による

（データアーカイブミニワークショップ@国立天文台 2020.1.29）

https://www.adc.nao.ac.jp/people/~furusawa/work/archive_ws/sekai_no_archive.pdf

1）Astronomical Data Archives Meeting @シドニー (2019.8.5-8.8)

- データアーカイブの運用・開発者が集まり、システムや技術的な課題を議論する会
- 参加アーカイブ関係者：CDS，MAST，IPAC，AAO，JHU，ESO，…
- 会議の様子
  - IVOAメンバーが中心の会議だったためVO・UI関係者が多く，各論の話が中心になった
  - クラウドへのデータ配置・サービス移行の検討は世界的に課題になっている…
    が，コスト面で有効かについては懐疑的であり，天文と企業側の意向にもギャップがある
  - DBやテーブルの分散化などDBの性能向上技術への関心が高まっている

各機関の報告（１）

- CDS
  - VOベースで世界のカタログ集約：TAPクエリ・ネームリゾルバを充実
  - HiPSを使ってほしい
- NASA IPAC
  - NASAがfundする基幹アーカイブ拠点の一つ。IRSA、NED、KOAなど
  - Keckアーカイブ（KOA）は運用がたいへん
    - 夜は観測者のもの：データがアーカイブを前提に取られていない
      - 望遠鏡が割り当てる名前からアーカイブ側のスキームへの変換など
- AAO(Australian Astronomical Optics Consortium) in 2018
  - AAL, AAO-Stromlo, AAO-マッコリー, AAO-シドニー大への改組
    - オーストラリアの全データの運用
    - ESOのパイプライン開発
  - ASVOを通じた活動: Data Central，CSIRO
  - AAL (~25% -> Data+Computing: $3M, 1FTE)
  - Data Central: AAT archiveーデータと解析コードを繋ぐこと・セキュリティが課題

各機関の報告（２）

- INAF
  - VO、欧州オープンサイエンスクラウドでの相互運用を目指す
- STScI MAST
  - NASAがfundする基幹アーカイブの一つ。HST、WFIRST、PanSTARRSなど
  - サイエンスプラットフォーム（ユーザ作業環境）構築ークラウド環境も試験
  - VO対応。MASTはNVOのテストベンチとして機能したがアーカイブ本体の維持と別のプロジェクトとしてお金が付く（そして止まる）のでたいへん。
  - 重要なアーカイブごとにgrantが付く
  - Common Archival Observation Model (CAOM by CADC)によるメタ情報フォーマットの統一化をHSTデータで試験 ←すべてのミッションでがうまくいくわけではない
- JHU SciServer
  - HPC環境，クラウドの利用が課題

各機関の報告（３）

- ESO
  - 科学データアーカイブの整備（public/ large programは必須）
  - データへのメタデータの整理・価値付加(curation)が重要
  - データの近くで解析・性能の最適化が課題
  - データコピーを持つ：年間生データ量～1PB，2-3年で2倍になる
- NAOJ
  - 複数プロジェクトのための独立な構成の計算資源を管理（レンタルシステムなど）
    - 有効なコストダウンが難しい
  - データサイズやデータ解析が当初の設計から進化し、システム・DB設計が難しい
  - バックアップが難しい
  - 既存DB技術の限界
  - プロジェクトの多重化による慢性的なマンパワー不足

まとめ

- 増加するファイル量（サイズ，個数）への対応の苦慮は共通．
  場所柄SKAの話は多かったが．
  - 運用的には保管場所，技術的にはDBへの登録速度，解析効率など
- クラウドへのデータ配置・サービス移行の検討は世界的に課題に
  なっている
  - が，全体としてコスト面や運用面で有効かについては懐疑的で，
    天文と企業側の意向にはギャップがある
- 分散化などDBの性能向上技術・ユーザ解析環境への関心が高まっている
- 参加者による偏りもあるが，データ保全や健全性確保よりは
  科学利用促進へのウエイトが高かった
  - 逆にアーカイブ従事者がそれに専念できる環境・役割定義があるのかも？
- NASA関連など寿命のあるプロジェクトへの投資で物事が動いている印象はあったが，全体として予算が足りないという話はあまり聞かなかった

2）国外可視・光赤外アーカイブの状況非公式調査

- 世界の主要な可視・光赤外の観測データを保持するデータアーカイブのうち，研究上のコネクションがある方を中心に，データアーカイブ運用を取り巻く状況について非公式のアンケートで尋ねた
- 目的
  - 主要データアーカイブの運用・開発のコアにいる人が何を大事にして作業に従事し，彼らのコミュニティにおけるデータアーカイブの位置づけをどのように考えているのかを知ることで，日本のアーカイブの在り方を考える材料とする
- 注意点
  - コンプリートな調査ではない
  - あくまで個人的な意見で組織を代表した意見ではない

機関リスト

- ご協力・お返事をいただいた機関
  - CADC（カナダ）
  - IPAC-IRSA/NASA（アメリカ）
  - ESA（欧州）
- お返事待ちの機関
  - ESO（欧州）
  - STScI-MAST/NASA（アメリカ）
  - AAO・Data Central（オーストラリア）

質問内容

- Q1. 何のデータを管理していますか。それはどのように決めていますか。
- Q2. 何人規模でどのような役割のメンバーで運用していますか。予算規模は。
- Q3. 観測所との役割分担はどうしていますか。
- Q4. データの利用頻度や科学的な価値などによって優先付けをしていますか？
  生データと処理済みデータの間ではどうですか。
- Q5. 計算機リソースの管理や調達はどうしていますか。
- Q6. あなたのデータアーカイブの一番重要な役割はなんですか。
- Q7. 予算措置は、その目的の実現のために適切ですか。
- 　困難がある場合、どのように対応していますか。
- Q8. コミュニティや所属機関の理解とサポートは得られていますか。

Q1. データの種類と決め方

- CADC
  - カナダの観測所群の全データ（新旧）＋HSTやJWSTなど＋観測所から頼まれたデータ
    - 2015までGeminiのデータも提供していたが観測所管理に移行
  - 望遠鏡によっては処理済みデータも提供
  - 保持出来る限り優先度は付けない．データごとの違いがないことがシステム設計上も有効．
- IPAC
  - NASAのIR・サブミリデータのほとんど（除 HST,JWST,WFIRST)＋
一部の米国・国際データ＋all sky surveys
  - 個人研究者の処理済みデータ（査読論文必要）
  - 大きなデータの決定は4-5年ごとのNASAの方針＆グラントの審査による。
  - 小さなデータはコミュニティの意見を聞いて。
- ESA
  - ESAミッションのデータ：ESACが全ミッションを支援する必要がある
  - 生データは一部、処理済みデータは必須

Q2. 何人で運用している？予算は？

- CADC
  - 20人フルタイム：開発8、運用4（不足）、研究者8（CADC全体の職員は120）
  - 研究者は１～数個のアーカイブの管理・デザイン、開発・運用もする
- IPAC-IRSA
  - 12人フルタイム等価：開発者~8、研究者~4（ただし、IPACデータサービス全体は150）
  - 予算の1/7が物品購入費
- ESA
  - 全体で30人規模
  - ミッションごとのアーカイブ科学者(ミッションチーム員）は1～数人

Q3. 観測所/データ提供者との役割分担

- CADC
  - アーカイブ側：データに対するテスト、テストの修正。ファイルは変更しない。
  - データ提供側：データの修正（アーカイブを無償提供しているから良く対応してくれる）。
  - Common Archive Operation Modelでメタ情報形式の共通化をしている。
  - Geminiとの連携では彼らの希望を元に多くのインターフェースを策定した。
- IPAC
  - アーカイブ側：メタ情報がアーカイブの規約に沿っているかのテスト、
    データが満たすべき条件のドキュメンテーション。ファイルの修正ではなく利用価値を付加する。
  - データ提供側：正しい/validatedデータの提供、科学的なデータ評価。
  - パイプライン策定時から密に連携。
  - KOAではより多くの連携作業があるだろう。
- ESA
  - 装置チーム：装置チーム自前のパイプラインを走らせ、データをアーカイブに提供
  - ESAC：　パイプラインを走らせる（大きなミッションではコンソーシアムと共同作業）
  - アーカイブ科学者：ミッションチーム員、開発者とともにアーカイブの仕様・I/F策定・実装。
    ミッションによってはvalidationも行う。

Q4.データの優先度、生・処理済みデータ

- CADC
  - データの差別なし。同じシステムで等価に扱える。
  - 優先度は処理済みデータ：よりユーザに有用だから
  - アーカイブ種類によっては解析もする（CFHT）
  - 利用頻度の低いデータ・古いデータもあるが、大した量ではない。
    すべて同等に保持する方が複数のシステムを組むよりシンプルでコストも低い。
- IPAC
  - 優先付けはNASA、コミュニティからの入力、レビューによる。
  - 生・処理済み両方を保持。ほとんどのミッションについて生データを永久保持。
  - 処理済みの方が利用者は多い。生データは解析手法を変えた将来の仕事などに必要。
- ESA
  - 処理済みデータを優先。
  - ミッションによる。

Q5. チーム人員で計算機管理しているか

- CADC
  - 部分的に。
  - ストレージはCompute Canada（リサーチクラウド）に移動。
    容量は稼げるが必ずしも応答が良くない。ネットワークはCANFER。
  - 全コピーはインハウスで持っている
- IPAC
  - IPACの計算機管理グループによる。IRSAも協力。12PBストレージを保持。
- ESA
  - 地上班のインフラを担当する別チームによる管理。

Q6. 貴アーカイブの重要な役割とは

- CADC
  - 1. 全カナダ天文学データの集約と利用提供
  - 2. データ保存のストレージを提供
  - （処理済みデータ提供、計算環境提供、データ解析支援）
- IPAC
  - 新しいデータ投入
  - 替えの効かないIRデータアーカイブの維持
  - 先端研究の支援
- ESA
  - 信頼のおけるデータ。ミッションエキスパートによる検証付き。
  - 長期のデータ保持。
  - 科学指向で使い勝手の良いサービス。

Q7. 予算は適切か

- CADC
  - 適切
- IPAC-IRSA
  - 適切
  - 特別な活動については追加予算を探す（定常運用部分は適切）
- ESA
  - 各ミッションが予算措置し、フェーズにより不足もあり。
  - レガシーデータアーカイブのための予算を模索（Herschel、Plank）

Q8.コミュニティ・機関の理解とサポート

- CADC
  - 得られている。
  - Long Range プラン(10年計画）でビッグデータを扱う拠点として位置づけられている
  - 直接の所属組織（Herzberg A&A）からの評価は適切、その上のNRCの評価は必ずしも盤石ではない
- IPAC
  - 得られている。
  - IPAC・NASA双方、アーカイブ科学を重視。コミュニティも同様。
- ESA
  - 年々、重要性への理解が高まっている。
  - 近年の論文はほとんどがアーカイブデータから生まれている。

他コメント

- CADC
  - 半分以上の論文はアーカイブから生まれ、
    アーカイブは観測所のインパクトを2倍にする。コスト有効。
  - アーカイブI/F利用とデータ利用のバリアを下げるのが重要
- IPAC-IRSA
  - IRSA以外にNED、Exoplanet Archive、KOAもある。NEDとExoplanet Archiveは処理済みデータのみ
  - NASAからの予算付け：Astrophysics Data CurationとArchival Research Programを通している。
  - アーカイブは査読論文生産を倍増している（HST、Spitzerは半分以上アーカイブから）のでNASAは予算を付ける

まとめ

- 活動維持のための予算措置は適切
- コミュニティとしてデータアーカイブを維持するための組織立て・予算の仕組みがある
- 過去の生データを保持することには現時点では問題がない
- データ提供者がデータをvalidateし、データアーカイブはテストしロードし、解析のための情報付加（curation）に注力する（それが出来る）
  - ただし生データやデータ品質をどの程度きちんと整理・評価されているのかは、より装置・ミッションに近い側を調査しないとわからない
- 意識的には処理済みデータの優先度が高い
- アーカイブデータによる論文生産率が高いことを自負